Chapter 90
An Insight into State-of-the-Art Techniques for Big Data Classification

Neha Bansal
Department of IT, Indira Gandhi Delhi Technical University for Women, Delhi, India

R.K. Singh
https://orcid.org/0000-0001-8729-2293
Department of IT, Indira Gandhi Delhi Technical University for Women, Delhi, India

Arun Sharma
Department of IT, Indira Gandhi Delhi Technical University for Women, Delhi, India

ABSTRACT

This article describes how classification algorithms have emerged as strong meta-learning techniques to accurately and efficiently analyze the masses of data generated from the widespread use of internet and other sources. In particular, there is need of some mechanism which classifies unstructured data into some organized form. Classification techniques over big transactional database may provide required data to the users from large datasets in a more simplified way. With the intention of organizing and clearly representing the current state of classification algorithms for big data, present paper discusses various concepts and algorithms, and also an exhaustive review of existing classification algorithms over big data classification frameworks and other novel frameworks. The paper provides a comprehensive comparison, both from a theoretical as well as an empirical perspective. The effectiveness of the candidate classification algorithms is measured through a number of performance metrics such as implementation technique, data source validation, and scalability etc.

DOI: 10.4018/978-1-7998-2460-2.ch090

1. INTRODUCTION

In near future, the data figure will double at least other two years. Few examples that show this massive and fast increase are: Google processes data of 100 PB, Facebook generates data of 600 TB, Baidu, a Chinese company, processes data of 10-100 PB, and Taobao, a subsidiary of Alibaba, generates data of around 10 TB per day from online transactions (Jin, Wah, Cheng, & Wang, 2015). As per reports of IDC, marketing of big data will reach to $32.4 billion by 2017 (International Data Corporation [IDC], 2013). Big datasets typically include masses of structured and unstructured data that require analysis to be performed in real-time. Accessing data, domain knowledge of data, privacy of data, computing and mining data are the major challenges of big data as identified by Wu et al. (2014).

Big data analytics will require a redesigning of existing data mining algorithms and execution in parallel frameworks. Among many alternatives, the MapReduce model and its distributed file system (Zhao, & Pjesivac-Grbovic, 2009; Dean & Ghemawat, 2008) offers a scalable and fault tolerant framework to address big datasets analysis. Hadoop has been the most relevant implementation for MapReduce (Lam, 2009) over the last few years. However, despite its various good properties, Hadoop-MapReduce framework has some drawbacks like the insufficient handling of iterative jobs and slow performance when combining data from multiple sources. Several alternatives like Spark (Zaharia, Chowdhury, Franklin, Shenker, & Stoica, 2010) and Flink (Alexandrov et al., 2014) are looked upon as the new standards.

Pre-processing followed by an appropriate classification technique is one of the most significant approaches in data mining. Classification is the process of classifying data into structured classes or groups. A variety of classification models may be constructed such as fuzzy logic, classification rules, support vector machine, neural networks and fuzzy neural etc. For big data classification, these traditional algorithms need to be modified using various ways. A comprehensive review of scalable machine learning algorithms for big data has been done by Gupta et al. (2016). The authors concluded that most of the previous works have focused on making traditional serial techniques scalable. New techniques build specifically for big data are not worked upon much. The present paper discusses various classification approaches for big data sets. The paper is well organized into five different sections. Section 1 outlines with general introduction of big data and big data classification techniques. Section 2 explores various problems of big data classification. Section 3 elaborates the employed research methodology emphasizing as on what basis the research papers have been selected and also the criteria for their inclusion in our work. It also presents the various research questions to be addressed by our research. Sections 4 and 5 provide a detailed survey and comparative analysis of various classification approaches for big data. Sections 6 and 7 contain concluding remarks and future scope.

2. BIG DATA CLASSIFICATION ISSUES AND CHALLENGES

Different classification approaches like Fuzzy Logic, SVM, k-NN, Neural Network and hybrid of approaches have been used as traditional methods of data mining. Fuzzy logic is a computing technique that is based on the degree of truth not like the crisp systems based on true & false (1 or 0) techniques (Wu et al., 2014; Sharma & Padamvar, 2013). k-Nearest Neighbor (k-NN) is based on the theory of linear interpolation. It judges the class of an object by examining its K nearest neighbor's class labels. However, for big data, the algorithm has to be modified as time and cost that will be required for comparing the distances between existing objects and the newcomer would be unacceptable. Keller et al.